White Paper for Tools for Listening to Text-in-Performance
By Marit MacArthur and Neil Verma

Digital Humanities Advancement Grant
National Endowment for the Humanities
Application Number: HAA-258799-18
Project Directors: Neil Verma (Northwestern University) and Marit MacArthur (CS Bakersfield and UC Davis)
Project Dates: 01/01/18 - 09/30/19
Institution: Northwestern University

Overview

This project developed, integrated, disseminated and supported humanistic research using two state-of-the-art, open-source, user-friendly tools—Gentle and Drift—for analysis of recorded texts-in-performance. The principal fields of the project are literary, media, performance and film studies.

The project was organized and pursued by Project Directors Neil Verma (Northwestern) and Marit MacArthur (CS Bakersfield and UC Davis), Co-Investigator Mara Mills (NYU) as well as consultants Lee Miller (UC Davis) and software developer Robert Ochshorn. There were seven members of an advisory board comprised of senior scholars in DH, sound studies and voice studies, as well as 20 user-testers from 11 states across the U.S. and Canada who worked with their own recordings of texts-in-performance independently. All participants in the project had online training in how to use the tools, held online meetings to troubleshoot and trade best practices, completed two surveys about their use, and met in person twice at meetings of the Great Lakes Association for Sound Studies (GLASS) at the University of Wisconsin, Madison and Northwestern University, both in 2018.

At the end of the project, dozens of papers and presentations had been prepared based on this research, and a new version of the key digital tool Drift had also been developed, based on feedback from humanities researchers describing their needs.

Analyzing Text-in-Performance

How quickly someone speaks, how often and how long they pause, how they vary their pitch—all of these factors influence listeners' perceptions of a speaker, and what makes that speaker engaging, boring, dramatic, monotone, and so on. These factors can matter as much as, if not more than, the verbal content of a statement—imagine listening to someone insist, *I'm not angry!* in an angry tone of voice.

Tens of thousands of hours of recorded poems, radio plays, talking books, political speeches, sermons, podcasts and other cultural and literary texts-in-performance have long awaited the serious attention of digital humanities scholarship. Some of the main datasets for this project

included, for example, 324 of Orson Welles's radio plays at the Lilly Library at Indiana University, the 75-year Talking Book Collection of the American Federation of the Blind, and the University of Pennsylvania's Pennsound, with spoken audio from 564 poets. Each of these datasets show text coming to life when they are spoken and heard in recorded performances. Yet many humanists have lacked methods for analyzing the non-verbal aspects of performance in ways that engage with social, aesthetic and cultural contexts. And the archive of possible research is expanding thanks to recent initiatives such as the Radio Preservation Task Force at the Library of Congress, and the SpokenWeb initiative surrounding performed literary texts, which are making these recordings more available than ever. In light of this, tools are needed to help scholars build new practices of "close listening" that identify nuances of speech and maximizes what we can learn from recordings.

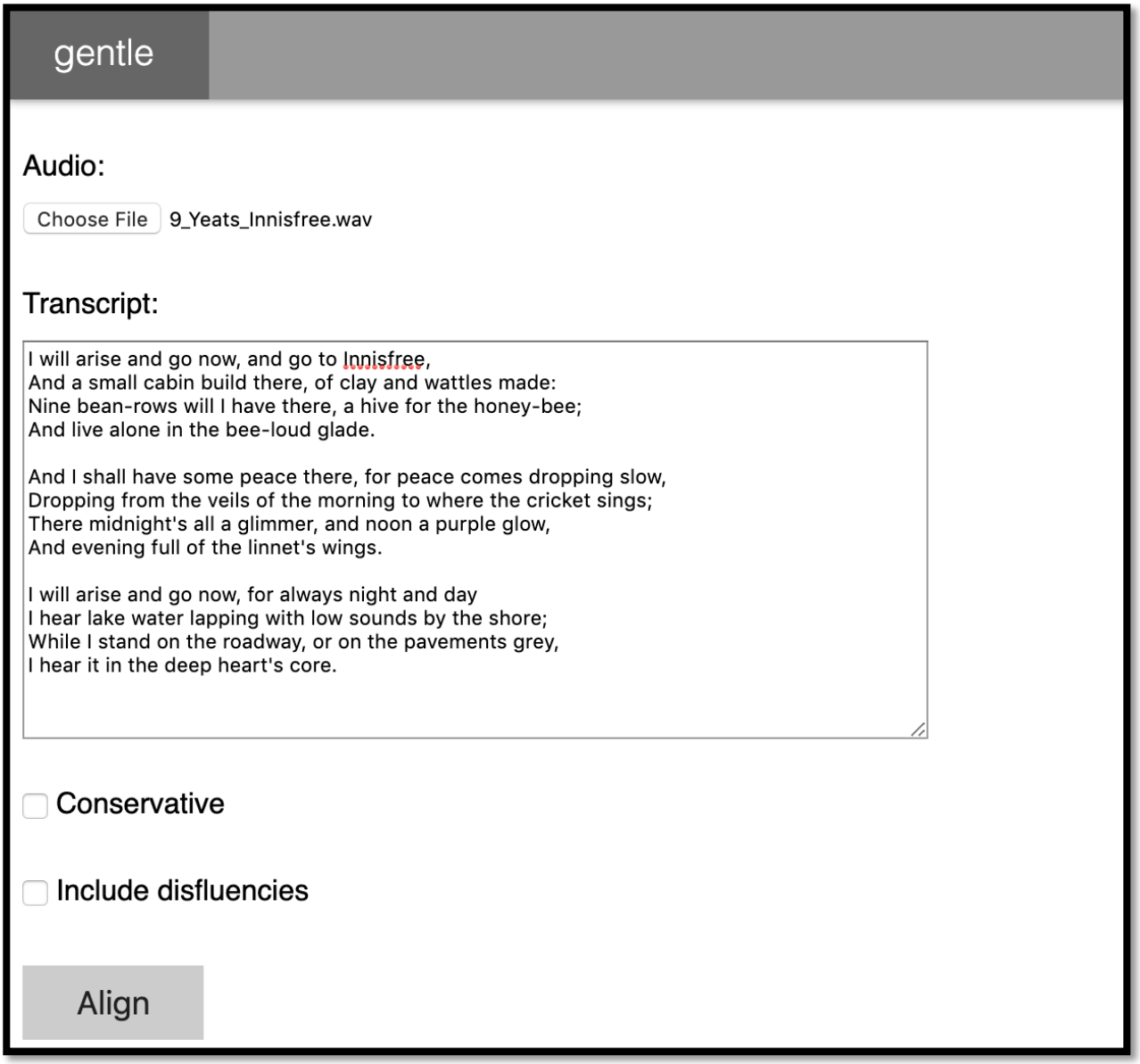
To rigorously practice close listening and empirically test received narratives about texts-in-performance and the history and evolution of vocal performance styles, "Tools for Listening to Text-in-Performance" has helped develop and disseminate robust, accessible tools that work well with low-quality, noisy audio, and that respond to the specific needs of humanist researchers with leading edge methods in speech recognition and signal processing. The project also made progress toward developing tools that facilitate and, when possible, automate the analysis of far more recordings than individual scholars could listen to in a lifetime. They can also help check the biases of human listeners and reveal performer choices: a recording may seem "monotone" when it really shows great variation of pitch; tools can help show vocal stereotypes associated with race and gender across performances in a given time period or medium; when a performer "mimics" another, digital tools can help tell us what aspects of the original speaker are being emulated, and what aspects are discarded.

With a multidisciplinary team of three principle researchers, the project began with prototypes of two tools, Gentle and Drift.

What are the tools?

Gentle is a forced aligner, which takes a recording of speech, aligns a transcript of speech with the recording, and returns precise timing duration for each word and pause. Gentle works well on noisy recordings because it is built on top of a state-of-the-art open-source speech recognition toolkit developed (with support from a National Science Foundation grant) at Johns Hopkins University, Kaldi, which uses modern neural network-based acoustic modeling, and which was trained on thousands of hours of telephone conversations.

Timing data can be used to characterize speaking rate, pacing, tempo, rhythmic patterns, and the degree of rhythmic regularity. Pause data can also be used to investigate whether, for instance, a poet pauses at line breaks, or whether and why an actor might be slowing down or speeding up the delivery of lines during a key monologue. The interface is very easy to use; a user simply uploads a file and pastes in a transcript. Choosing to "include disfluencies" will include vocalizations that are not words, like "uh" and "um"; the "conservative" setting will exclude them:



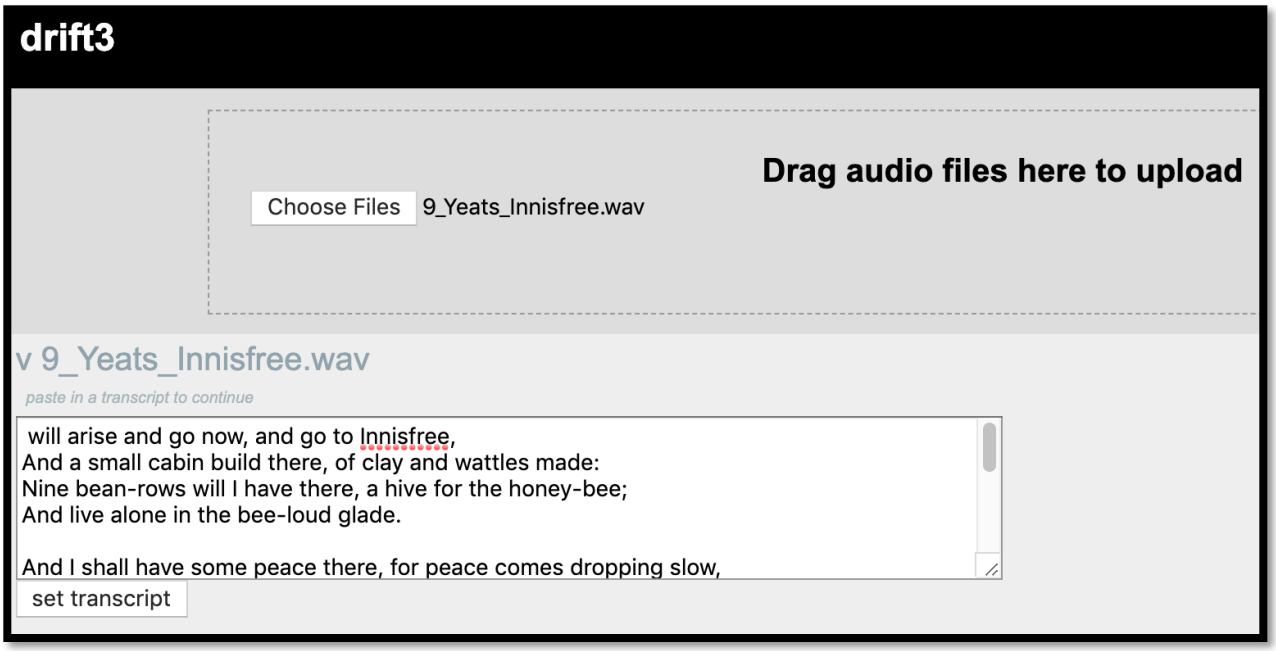
Here is an example of Gentle's timing data from a recording of a 1936 recording of William Butler Yeats reading "The Lake Isle of Innisfree" from PennSound, the online poetry audio archive at the University of Pennsylvania. Two members of the project team, Marit MacArthur and Lee M. Miller, discuss this recording in more detail in a multimedia piece, "After Scansion: Visualizing, Deforming and Listening to Poetic Prosody.") We can see that the longest pause, of .48 seconds, occurs between the second "go" and "to," before "Innisfree," which is not where a line break occurs. The pause serves to emphasize his destination. "Innisfree" and "wattles" are marked as "unk," or "unknown, because they are not in Gentle's lexicon; in such cases, the pause length can be calculated manually.

A	B	C	D	E
Word	Word	Start	End	Pause Duration
I	i	0.23	0.37	0
WILL	will	0.37	0.57	0
aRISE	arise	0.57	1.26	0
and	and	1.26	1.53	0
GO	go	1.53	1.84	0
now	now	1.87	2.11	0.03
and	and	2.13	2.17	0.02
GO	go	2.65	2.88	0.48
to	to	2.9	2.92	0.02
INnisfree	<unk>	2.92	4.2	0
And	and	4.585	4.74	0.385
a	a	4.74	4.89	0
SMALL	small	4.89	5.74	0
CAbin	cabin	5.81	6.41	0.07
BUILD	build	6.42	6.91	0.01
there	there	6.91	7.31	0
of	of	7.32	7.41	0.01
CLAY	clay	7.56	8	0.15
and	and	8	8.03	0
WAttles	<unk>	8.1	8.93	0.07
MADE	made	8.96	9.18	0.03

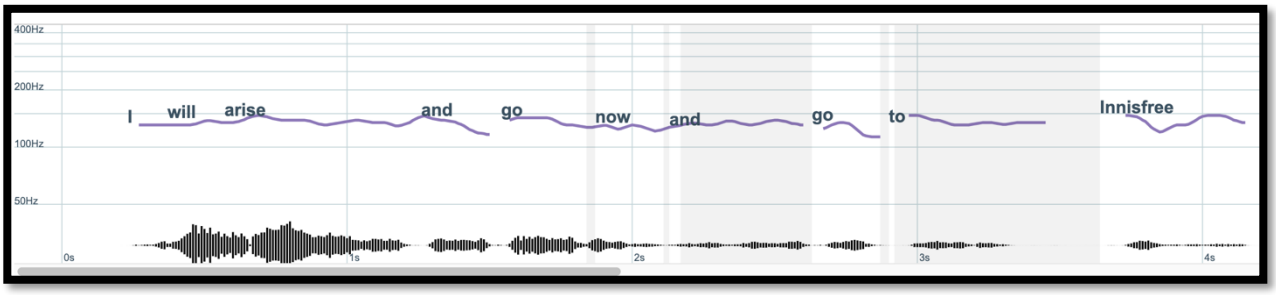
We can also calculate the speaking rate, which is a fairly slow 140 Words Per Minute, quite typical of the pace of public poetry readings. This is contrast to conversational speech, as two members of the project team explained in a study of sample recordings of 100 American poets, "Beyond Poet Voice: Sampling the (Non-)Performance Styles of 100 American Poets."

Drift, is a highly accurate pitch-tracker that also incorporates the forced alignment features of Gentle, visualizing an intonation pattern, or pitch trace, over time and aligning it with a transcript. Drift uses an algorithm developed by Byung Suk Lee and Daniel P. W. Ellis at Columbia University to work with precise accuracy on the noisy, low-quality vocal recordings common in the audio archive, Drift measures what human listeners perceive as vocal pitch (the fundamental frequency, the vibration of the vocal cords, as measured in hertz) every 10 milliseconds in a given recording. Drift was prototyped in 2016 by Ochshorn and Hawkins with support from Marit MacArthur's ACLS Digital Innovations Fellowship.

Like Gentle, Drift is easy to use. This is the interface; the user simply uploads a recording and pastes in a transcript:



Here is Drift visualizing the intonation of the opening of “The Lake Isle of Innisfree”:



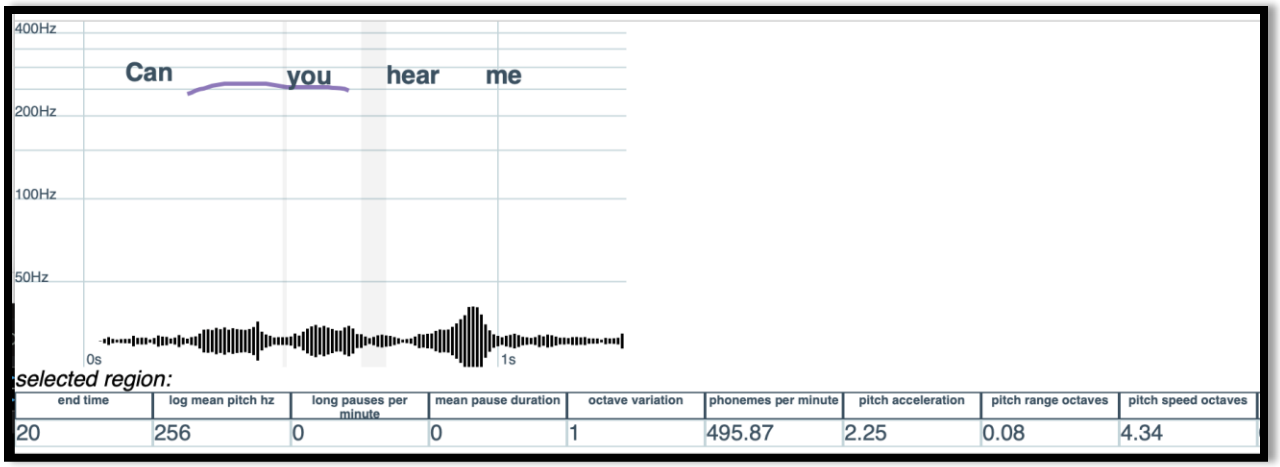
As Yeats sounds like he is using a near incantatory monotone, the pitch remains fairly flat. The data provided by Drift supports that impression, as the pitch values are within a narrow range:

0.34	130.86	I
0.35	130.86	I
0.36	130.86	I
0.37	130.86	I
0.38	130.86	will
0.39	130.86	will
0.4	130.86	will
0.41	130.86	will
0.42	130.86	will
0.43	130.86	will
0.44	130.86	will
0.45	130.86	will
0.46	130.86	will
0.47	130.86	will
0.48	130.86	will
0.49	130.86	will
0.5	134.7	will
0.51	134.7	will
0.52	138.64	will
0.53	138.64	will
0.54	138.64	will
0.55	138.64	will
0.56	134.7	will
0.57	134.7	will
0.58	134.7	arise
0.59	134.7	arise
0.6	134.7	arise
0.61	134.7	arise

In other cases, these tools can be used to test subjective impressions. For instance, consider Jonathan Mitchell's 2012 radio play "Tape Delay." The play surrounds a phone conversation before a date between Ben and Erica. Erica calls Ben while he is waiting for her at a bar, but his responses to her statements drive her away, and she refuses to meet him. Ben is perplexed, until he replays a tape of the conversation and hears that his voice sounds boorish from her end. We are led to believe that we are hearing the very same thing we first heard, only from another perspective, but with so much noise in the background, it's hard to tell. Using Drift shows that there are in fact two entirely different versions of the call in the piece. Ben uptalks in the first version, as he is trying to be charming, his voice rising from around 160 to 250 Hz. But when we hear him on a tape of the same call later on, the pitch shows almost no curve at all, and he speaks almost ¼ speed, which makes him sound annoyed and arrogant.

selected region:

end time	log mean pitch hz	long pauses per minute	mean pause duration	octave variation	phonemes per minute	pitch acceleration	pitch range octaves
20	180.04	0	0	5	909.09	1.85	0.54



selected region:

end time	log mean pitch hz	long pauses per minute	mean pause duration	octave variation	phonemes per minute	pitch acceleration	pitch range octaves	pitch speed octaves
20	256	0	0	1	495.87	2.25	0.08	4.34

In this case, Drift and Gentle help settle a difficult interpretive problem about the very nature of the text itself. By settling the problem, the events of the remainder of play take on an entirely new meaning, and we get a glimpse of some important slight-of-hand on the part of producer Jonathan Mitchell.

How were the tools applied?

To demonstrate and develop these tools, we applied them to four key datasets: an archive of 324 recordings of Orson Welles's radio plays, recently digitized by the Lilly Library at Indiana University with a National Recording Preservation Foundation grant, which co-director Neil Verma is studying; the entire 75-year Talking Book Collection from the American Federation of the Blind, the forerunners of today's audio books, which feature trained radio and stage actors, digitized with a National Science Foundation grant, which co-investigator Mara Mills is studying; and two archives of recorded poetry, the University of Pennsylvania's PennSound (more than 5,000 hours of poetry audio by 564 poets) and selections from the Library of Congress Archive of Recorded Poetry and Literature (100 hours of poetry readings by 63 poets), which co-director Marit MacArthur is studying.

Over the grant period, our multidisciplinary team of 20 user-testers were trained to apply Gentle and Drift to the archives that they are studying, including talking books, early recordings of vaudeville actors, classic Hollywood cinema, different actors' readings of T.S. Eliot, TED talks, podcasts, etc. At two meetings of GLASS (Great Lakes Association for Sound Studies, presently the only sound studies association in the U.S.), at the University of Wisconsin, Madison, and Northwestern University, the project team met to present their research, and then completed surveys on their experience using Gentle and Drift. Their responses influenced the design of Drift3, which was released in February 2018. It is available for free download and installation here: https://rmozone.com/drift/. Gentle is required to run Drift, and it continues to be available for free download and installation as well: https://lowerquality.com/gentle. Links to Gentle and Drift will remain on the Tools for Listening to Text-in-Performance website at Northwestern, so that members of the public can download them easily:

https://textinperformance.soc.northwestern.edu/. The source code for the tools is also available on GitHub: https://github.com/lowerquality/gentle and https://github.com/strob/drift3.

Over the grant period, the project team has presented their work at many academic conferences, and are starting to publish it as well, including in a forthcoming special issue of *Sounding Out!* in 2020: https://soundstudiesblog.com/. Here is a list of presentations and forthcoming publications:

Marit MacArthur, co-director of the project, with Lee M. Miller, a consultant on the project, led workshops on Gentle and Drift at the SpokenWeb Sound Institute at Simon Fraser University in May 2019. MacArthur also gave a workshop on the tools at Performance Studies International at the University of Calgary in July 2019. Casey Long, a film studies PhD candidate UW Madison, presented "Hyper-Emphasized Dialects and Vocal Performance in Video Games: Cindy's Southern Accent in Final Fantasy XV," at the Society for Cinema & Media Studies Conference, in Seattle 2019, and The analysis of vocal performance in Classical Hollywood cinema: "Volleying" pitch, loudness and tempo in Lubitsch's DESIGN FOR LIVING (1933), at the Music and the Moving Image Conference, New York City, 2018. Adam Hammond, an assistant professor of English at the University of Toronto, and graduate student Jonathan Dick, now a PhD candidate in English at the University of Pennsylvania, presented "They Do the Police in Different Voices: Computational Analysis of Digitized Performances of T. S. Eliot's The Waste Land," at the Association for Computation in the Humanities, in Pittsburgh, PA, in July 2019, and "Performing the Literary Vernacular: a Computational Approach to Racialized Voice in Jean Toomer's 'Kabnis,'"at Digital Humanities 2019, in Utrecht, Netherlands, in June 2019. They also presented related work at the Canadian Society for Digital Humanities, in Vancouver, BC, in May 2019 and at the SpokenWeb Symposium, Vancouver, BC, May 2019. Adam Hammond and Jonathan Dick, "'The Mold That's Branded on My Soul': A Computational Approach to Racialized Voice in Jean Toomer's 'Kabnis,'" SpokenWeb Symposium, Vancouver, BC, May 2019. Their essay, "'The Mold That's Branded on My Soul': A Computational Approach to Racialized Voice in Jean Toomer's 'Kabnis,'" will appear in the Proceedings of the Digital Humanities 2019 Conference (June 2019): 1–4. Several members of the team at UW Madison, Jacob Mertens, Eric Hoyt, and Jeremy Wade Morris, presented "PodcastRE: Saving and Studying New Sounds," at the Association for Computers and the Humanities, Pittsburgh, July 2019. Their book chapter, "Drifting Vocal Performances: Studying Vocal Pitch and Frequency in Podcasting with Digital Tools," is forthcoming in Saving New Sounds: Dispatches from the Podcaster's Project, ed. Jeremy Wade Morris and Eric Hoyt (Ann Arbor: University of Michigan Press, 2020). Jacob Smith, associate professor of communications at Northwestern University, has written an essay, "The Courtships of Ada and Len: Mediated Musicals and Vocal Caricature Before the Cinema," which is forthcoming in the Oxford Handbook of Cinematic Listening. Chris Mustazza, a PhD candidate in English at the University of Pennsylvania and associate director of PennSound, presented "Listening at Distances" at Plotting Poetry: Bringing Deep Learning to Computational Poetry Analysis Conference at the Freie Universität Berlin in September 2018 and "In Search of the Sermonic: Hearing Sonic Genre in Poetry Recordings" at Plotting Poetry: Machiner La Poésie. Université de Lorraine. Nancy, France in September 2019.

What have the reactions been?

A unique feature of the Text-in-Performance project was the close way in which participants and users engaged with the developer, Robert Ochshorn. After seeing a demonstration of early versions of the tools at GLASS meetings in Winter, 2018, user-tester researchers working on everything from early Vaudeville to TED talks took an online course from MacArthur about how to use Gentle and Drift. Then participants were asked to fill out a survey on what worked and what they struggled with – as well as what additional features could be useful – and these were passed along to Robert Ochshorn, who used the surveys to inform further development of the tools culminating in the creation of Drift 3, which was debuted at a second GLASS conference in Fall 2018 and released in the Winter of 2019.

In this way, rather than simply adopting pre-existing tools for their work, humanist research directly informed the design, interface and capabilities of these tools based on practical experience. After the process was complete, a second survey was launched in Summer of 2019 asking user-testers to reflect on their experience of using the tools, once they had been refined based on prior feedback. Here are some of their statements.

On how Gentle proved useful for analysis:

> "It is irreplaceable. This is particularly true of presentations and conferences, where a visual demonstration of sound tools is helpful in aiding listeners to hear the nuances of vocal performance. It helps me in a different way-- to notice patterns or anomalies more quickly and to avoid using the more clunky linguistic software previously available."

> "It has allowed me to ask large-scale quantitative and statistical questions about materials that I have long been approaching in a small-scale qualitative close-reading or ethnomethodological way: namely what sorts of rhythmic habits characterize particular genres of technologically mediated speech and how can these habits be correlated to particular words?"

> "It's allowed me to quantify aspects of pacing that I'd previously been able to address only impressionistically, or to measure only with great effort (and to an accordingly limited extent). It has let me trace gradual changes in performance style, by comparing recordings made at different times, with greater confidence and with greater precision."

> "Gentle was excellent as a forced aligner, and allowed us to make rough transcripts of many podcasts we did not have transcripts for."

> "Prior to using Gentle, I don't think I was asking questions about pace and pauses between words. Now, it's something that I am investigating much more carefully. I'm also interested in studying a range of other issues like dynamism, tone, and various other elements related to delivery."

On how Drift proved useful for analysis:

> "Drift helped me significantly […] It sharpened how I understand speech and the voice, its details and peculiarities. […] It's also made me interrogate what I mean when using voice and sound terms, as I reached to describe what it was I heard."

> "Drift proved very useful for our research into the vocal performances of podcasters. On a micro scale, it let us see how pauses and pitch inflections were used to accomplish specific strategies."
>
> "Drift helped me to make the discovery that the practice of maintaining tension in the corners of the lips (a typically feminine practice) raises the pitch of a sibilance and of the following vowel above the practice of relaxing tension in the corners of the lips (a typically masculine practice)."
>
> "When it worked on my recordings, it gave me the opportunity to compare absolute values of a given recording with results from a specific fragment. That was very useful."
>
> "I am using [Drift] to understand whether particular genres can be identified by their prosodic signatures. It is helping me to understand what empirical facets of recordings make them sound sermonic and thus which might be used to reassess generic categorizations of modernist poetry."

On how Drift and Gentle have helped scholars reshape their research questions:

> "Computational audio analysis of pitch and timing data for readings of The Waste Land produces results that correspond well to many human listeners' subjective impressions (i.e., the tools seem to work) -- But these subjective impressions vary between individuals, and many individuals find their impressions difficult to articulate precisely. Thus computational analysis provides a different way of listening — one that may be able to notice details that human listeners can't — but, perhaps most usefully, provides a quantitative vocabulary against which to test and refine subjective impressions."
>
> "They have changed my research questions by expanding the sorts of questions I'm able to consider, namely giving a quantitative context to more interpretive projects."
>
> "The use of Gentle and Drift allowed me to analyze general trends in pitch range and pitch average across Jones and Spencer's enactment of various ethnic types. These tools also helped me to examine the finer grained dynamics of short examples of vocal performance, and thereby develop arguments about how the specific cadences and sonic tropes heard on these records might have aligned with the visual and thematic conventions of different stage types."
>
> "Gentle and Drift […] allowed students who have no prior experience in the analysis of performed speech to externalize what they were hearing and begin to develop a critical discourse for the analysis of spoken word. This was more successful and engaging for students than prior, non-mediated activities keyed to similar learning goals have been in the past."

What did we learn along the way?

There is a rich tension in the digital humanities between the necessarily small-scale, qualitative research that generates key insights, and the large-scale, more quantitative research that can test such insights and their broader applications. This project was no different. The more that our user-testers discovered insights about their archives using Drift and Gentle, describing the performance style of an individual actor, for instance, the more they were eager to study larger groups of recordings. To that end, the project team made steps toward incorporating a large-scale approach to pitch and timing patterns in performative speech (such as pitch range, rates of pitch change, length and frequency of pauses, and rhythmic predictability) called Voxit, which Miller and MacArthur had developed in previous research using Matlab and WORLD, a

state-of-the-art speech synthesis and analysis system developed in Japan: https://github.com/MillerLab-UCDavis/Voxit. Taking the pitch and timing data from Drift and Gentle, the project team was able to generate a script in Python that would calculate the same measures as Voxit, but without the batch uploading option allowed by Voxit. Progress was also made toward developing direct calculation of the Voxit measures in Drift.

Where do we go from here?

Our tools are helping to create new kinds of analysis that professors can use in the classroom, and a number of our team members already are, including Neil Verma at Northwestern University, Tanya Clement at the University of Texas at Austin, Kenneth Sherwood at Indiana University of Pennsylvania, and Chris Mustazza at the University of Pennsylvania. For example, vocal recordings of speeches, poems, films or dramas can be analyzed and sound patterns visualized with our tools, answering what makes some performances seem "charismatic" or "monotone," what vocal styles define gender or racial expressions, and how performers come to understand their own voices. Other researchers are using these tools to answer questions about where vocal stereotypes come from, or how certain voices come to seem “made for radio” while others do not, in particular social and cultural contexts.

The further development of Gentle and Drift, and of the Vox it measures within Drift and without the need for a Matlab license, is supported by SpokenWeb.ca, a 7-year Can$2.5 million SSHRC partnership grant (2018-2025). SpokenWeb is a project to “develop coordinated and collaborative approaches to literary historical study, digital development, and critical and pedagogical engagement with diverse collections of literary sound recordings from across Canada and beyond. In addition to teaching our multidisciplinary project team of approximately 30 scholars across the U.S. (including graduate students) to study recordings of voice in new ways, MacArthur and Miller demonstrated the tools to a group of 40 DH and poetry scholars, mostly Canadian, at the SpokenWeb Sound Institute in May 2019 at Simon Fraser University.

These scholars installed Gentle and Drift on their personal computers, and will form a new user-tester group moving forward, as part of the 7-year SpokenWeb SSHRC grant project. The open-source, user-friendly tools of Gentle and Drift have also been improved and refined to suit the needs of the diverse project team, including sharper visualization of pitch contours, the ability to play a recording and see a pitch contour at once, clearer alignment of text, scrolling features with overview and close-ups, the use of a logarithmic scale to allow for fair comparisons between speakers using higher and lower pitch ranges, and readier access to and analysis of pitch and timing data.

All of these refinements—and importantly, the fact that the tools are open-source, with online demos and source code on GitHub—have expanded the possibilities for studying poetry readings, radio plays, talking books, speeches, TED talks, NPR broadcasts, film acting, and early vaudeville acting. Five or ten years from now, the tools Drift and Gentle will be surely outmoded, but pitch and timing will always matter in performance styles, and the tools that come to build on Drift and Gentle will be less mysterious and more tailored to the needs of

humanists as a result of this project. And again, voice studies within the digital humanities is still quite novel – new publications have recently appeared from several University presses, and in 2016 a *Journal of Voice Studies* was launched – and the development and dissemination of Gentle and Drift with humanists' needs in mind, thanks to this NEH grant, has been a major step forward.